(ICW) The ADL is faking 36% to 43% of their recent antisemitism numbers

2025-08-26

By NRU

International Corruption Watch

Perilous times are coming

The International Corruption Watch (ICW) is a newly founded organization of independent journalists with the goal of bring to light the massive corruption that plagues governments and big tech worldwide. In an era where technology plays an increasingly central role in shaping public policy and opinion, we believe that the misuse of technology for corrupt purposes has become one of the most pressing issues of our time.

Follow us on social media:

BlueSky: https://bsky.app/profile/icw-nru.bsky.social

Substack: https://substack.com/@icwnru

Do you have any tips for a case of corruption in tech or government?

If so please send a message to **icw_nru@protonmail.com**

ICW is a self-funded operation. All donations are appreciated.

BTC: bc1qjy84xyr8u0majvtslzfz940hpkcp4hv3gjmms8

ETH: 0x0620C26611644Ec7287274713df0eD3Df93dBB82

Read our previous investigations!

(2024-11-17) The Youtube Alogirthm and Manufacturing Consent

(2024-11-30) What can we learn from the Andrew Tate data breach?

(2025-08-11) Meta Leaks Part 1: Israel & Meta, The Greatest Global Mass Campaign to Ever Exist

(2025-08-14) Meta Leaks Part 2: How To Kill A Social Movement

NOTE: from NRU

This story took more than 70+ hours from start to finish. We hope that the ADL spends this much time vetting the quality of their data, instead of spending their time seething over American college students or Miss Rachel.

This is our first, but not last story on the ADL.

*Article has been proofread by an LLM

Overview

We ran a validation experiment against the ADL's antisemitism dataset to verify its legitimacy.

We used LLMs to analyze the entirety of the ADLs antisemitism dataset, which contains over 33,000 entries. The data was analyzed in batches, with each batch including a **prompt** we **carefully crafted** to improve classification accuracy and reduce false negatives (i.e. incidents where we wrongly claim to not be antisemitic when in fact they are). The results of our investigation include both lower and upper bound estimates for the percentage of false positive entries (i.e. incidents that are not anti-semitic) in ADLs dataset (definitions of estimates found here). Below are the key findings of the investigation:

We determined that post October 7 2023, 36% to 43% of the incidents ADL claims are antisemitic are likely not.

Among the incidents we did not classify as antisemitic, the most common reasons for this conclusion were: "Conflating Anti-Zionism with Antisemitism" (7.78% to 9.93%), "Political Statements" (7.57% to 8.33%), "Lack of Explicit Antisemitic Content" (5.98%).

In contrast, prior to October 7 2023, we found only 11.34% to 16.13% to likely be false positive claims of antisemitism.

There is a sharp increase in false positive incidents submitted by the ADL following October 7 2023.

We also conducted a small validation experiment across various major LLM providers, and found that without initial prompting, all LLMs equate anti-Israel sentiment to antisemitism. This brings into question what kind of training data used or system prompt.

We discuss the role of the mainstream media in reporting on the ADLs antisemitism numbers (as well as their other claims on antisemitism), without verifying the integrity of the data.

Who is the ADL?

According to Wikipedia:

> *"The Anti-Defamation League (ADL), formerly known as the Anti-Defamation League of B'nai B'rith,[a] is a New York-based international non-governmental organization that was founded to combat antisemitism, as well as other forms of bigotry and discrimination.[3] The ADL is also known for its pro-Israel advocacy."*

What stands out from this introduction is that their "pro-Israel advocacy" is the second mentioned fact about the ADL. To anyone paying attention to the ADL for the last couple of years, this does not come as a surprise.

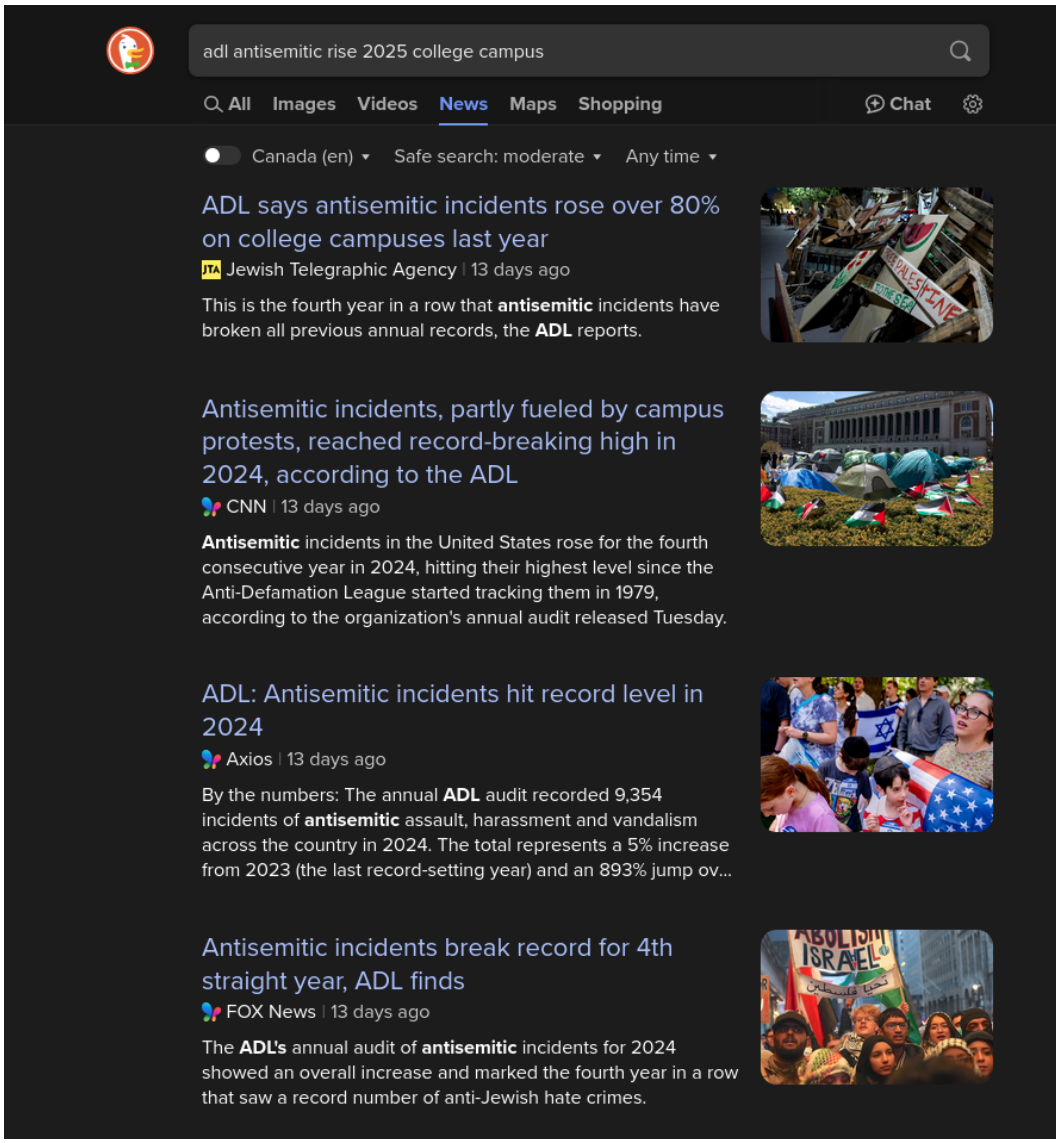
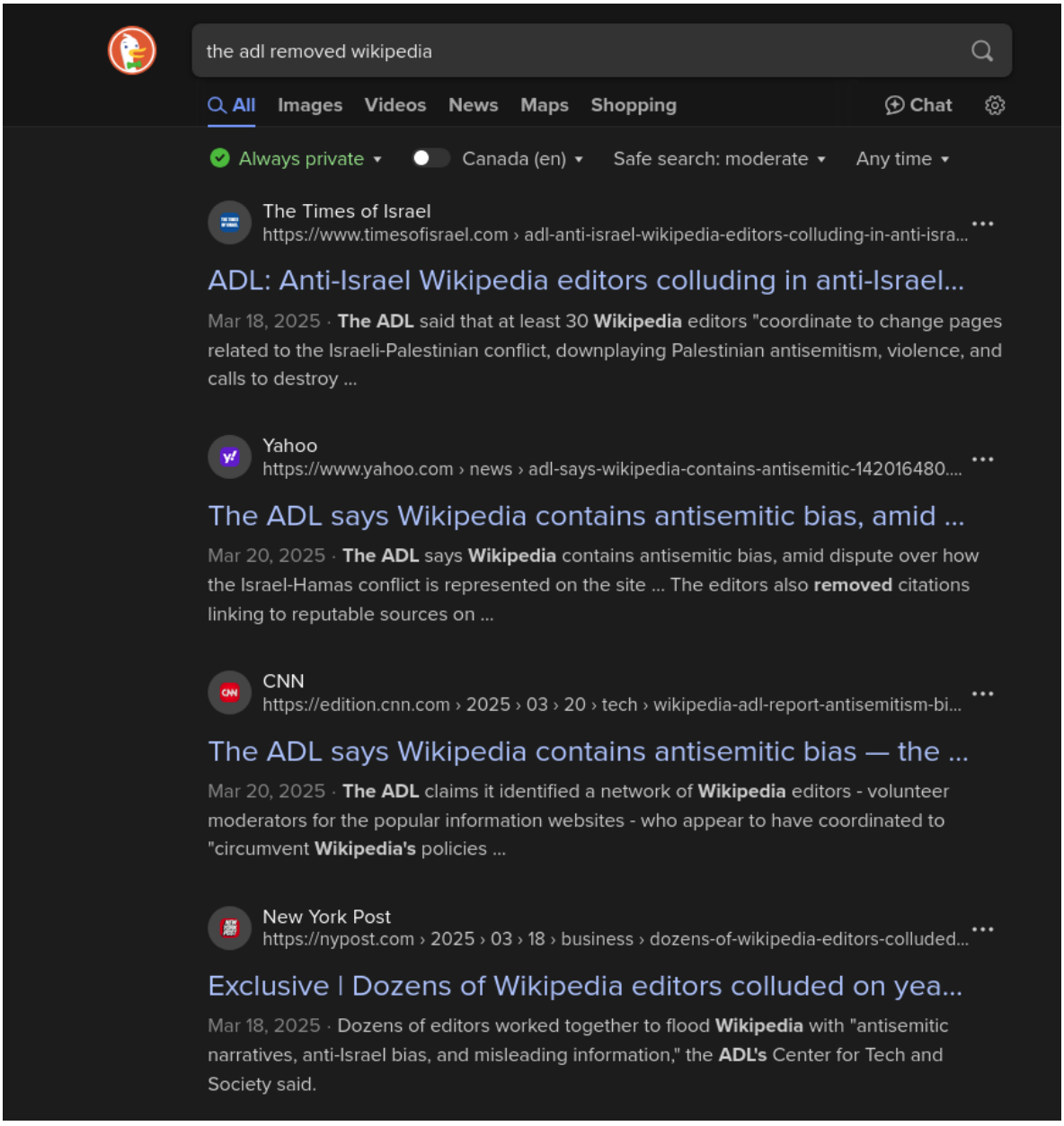

adl antisemitic rise 2025 college campus
All
Images
Videos
News
Maps
Shopping
Chat
Canada (en)
Safe search: moderate
Any time
ADL says antisemitic incidents rose over 80% on college campuses last year
Jewish Telegraphic Agency | 13 days ago
This is the fourth year in a row that antisemitic incidents have broken all previous annual records, the ADL reports.
Antisemitic incidents, partly fueled by campus protests, reached record-breaking high in 2024, according to the ADL
CNN | 13 days ago
Antisemitic incidents in the United States rose for the fourth consecutive year in 2024, hitting their highest level since the Anti-Defamation League started tracking them in 1979, according to the organization's annual audit released Tuesday.
ADL: Antisemitic incidents hit record level in 2024
Axios | 13 days ago
By the numbers: The annual ADL audit recorded 9,354 incidents of antisemitic assault, harassment and vandalism across the country in 2024. The total represents a 5% increase from 2023 (the last record-setting year) and an 893% jump ov...
Antisemitic incidents break record for 4th straight year, ADL finds
FOX News | 13 days ago
The ADL's annual audit of antisemitic incidents for 2024 showed an overall increase and marked the fourth year in a row that saw a record number of anti-Jewish hate crimes.

Figure 1. DuckDuckGo news search results for antisemitism of college campuses

Jonathan Greenblatt, the CEO of the ADL, has been hell-bent on eliminating all expressions of anti-Zionism, especially in the United States. The ADL has primarily focused its efforts on targeting American college campuses and social media apps.

In a leaked phone conversation, he mentions that Israel has a “major, major, major generational problem” among the younger generation, and this is a “TikTok problem, a Gen Z problem”. Just a few months later, pro-Israel lobbying groups managed to pressure the Biden administration into passing the “TikTok ban bill”. A bill that was going to ban TikTok was publicly promoted as a way to defend Americans against Chinese influence, but in the end, it turned out to be just another way to control the narrative on Israel.


the adl removed wikipedia
All
Images
Videos
News
Maps
Shopping
Chat
Always private
Canada (en)
Safe search: moderate
Any time
The Times of Israel
https://www.timesofisrael.com › adl-anti-israel-wikipedia-editors-colluding-in-anti-isra...
ADL: Anti-Israel Wikipedia editors colluding in anti-Israel...
Mar 18, 2025 · The ADL said that at least 30 Wikipedia editors "coordinate to change pages related to the Israeli-Palestinian conflict, downplaying Palestinian antisemitism, violence, and calls to destroy ...
Yahoo
https://www.yahoo.com › news › adl-says-wikipedia-contains-antisemitic-142016480....
The ADL says Wikipedia contains antisemitic bias, amid ...
Mar 20, 2025 · The ADL says Wikipedia contains antisemitic bias, amid dispute over how the Israel-Hamas conflict is represented on the site ... The editors also removed citations linking to reputable sources on ...
CNN
https://edition.cnn.com › 2025 › 03 › 20 › tech › wikipedia-adl-report-antisemitism-bi...
The ADL says Wikipedia contains antisemitic bias — the ...
Mar 20, 2025 · The ADL claims it identified a network of Wikipedia editors - volunteer moderators for the popular information websites - who appear to have coordinated to "circumvent Wikipedia's policies ...
New York Post
https://nypost.com › 2025 › 03 › 18 › business › dozens-of-wikipedia-editors-colluded...
Exclusive | Dozens of Wikipedia editors colluded on yea...
Mar 18, 2025 · Dozens of editors worked together to flood Wikipedia with "antisemitic narratives, anti-Israel bias, and misleading information," the ADL's Center for Tech and Society said.
Figure 2. DuckDuckGo search results for ADLs story on Wikipedia’s anti-Israel bias

Remember when Wikipedia stopped relying on the ADL as a credible resource on antisemitism? Well, just a couple of months later, dozens of media organizations wrote hit pieces against Wikipedia's editors, claiming they colluded in anti-Israel bias and "downplaying antisemitism." A couple of days later, the Trump administration threatened to remove Wikipedia's non-profit status. According to the article:

> *"The letter, which was first reported by The Free Press on Friday, brought up issues that have been raised recently by other organizations, including the Anti-Defamation League, which accused Wikipedia of being an online channel for "widespread antisemitic and anti-Israel bias" in March."*

Of course, when it comes to "downplaying antisemitism", the ADL and The Free Press have quite a bit of credibility on this topic.

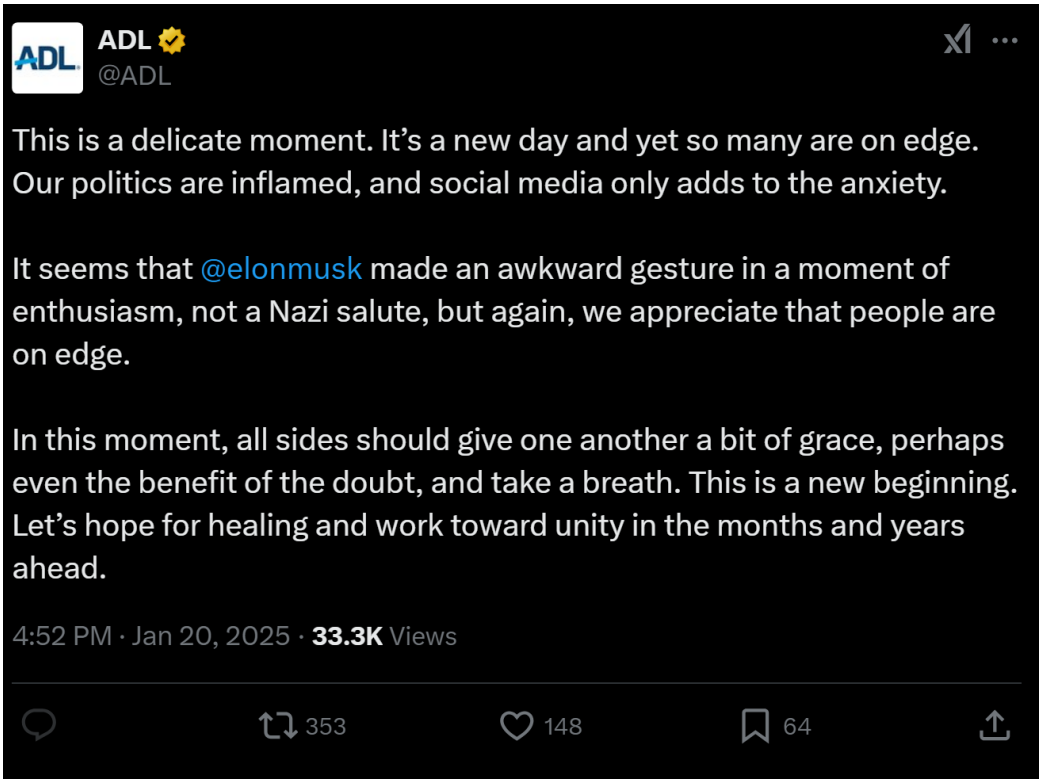

ADL
@ADL
This is a delicate moment. It's a new day and yet so many are on edge. Our politics are inflamed, and social media only adds to the anxiety.
It seems that @elonmusk made an awkward gesture in a moment of enthusiasm, not a Nazi salute, but again, we appreciate that people are on edge.
In this moment, all sides should give one another a bit of grace, perhaps even the benefit of the doubt, and take a breath. This is a new beginning. Let's hope for healing and work toward unity in the months and years ahead.
4:52 PM · Jan 20, 2025 · 33.3K Views
353
148
64

Figure 3. Tweet from the ADL defending Elon Musk Nazi salute

Bari Weiss
@bariweiss
Follow
"These stiff-armed salutes are not expressions of sincere Nazism but of an oppositional culture that, like a rebel band that keeps wearing fatigues after victory, has failed to realize it's no longer in the opposition."
Read @RichardHanania in @TheFP:
CPAC
The Guardia
Richard Hanania: I Can Explain Why the Nazi Salute Is Back
From thefp.com
3:23 PM · Feb 24, 2025 · 166.1K Views
Figure 4. Tweet from Bari Weiss of The Free Press defending Steven Bannon Nazi salute

Ironically, in this investigation, we provide strong evidence that the ADL is engaging in mass data manipulation to “overplay antisemitism”, likely to put more pressure on public and private organizations to censor anti-Zionist speech.

In this investigation, we show that after October 7, 2023, the ADL faked 36% to 43% of its antisemitism incident numbers. The fact that almost all major media outlets ran with the ADL's antisemitism reports without verifying the integrity of their data is even more troubling.

Strategy

The objective of this investigation is to assess the validity of the ADL's antisemitism data. When analyzing the data, we will aim to optimize for accuracy while reducing false negatives. In other words, we will be as generous or conservative as possible regarding the ADL's interpretation of "antisemitism" when handling ambiguous or borderline cases.

We are aware that any method we use for this investigation will still contain both false positives and negatives. This is why we will provide both lower and upper bound estimates of the ADL's claims. The definitions of these estimates are found in the following section.

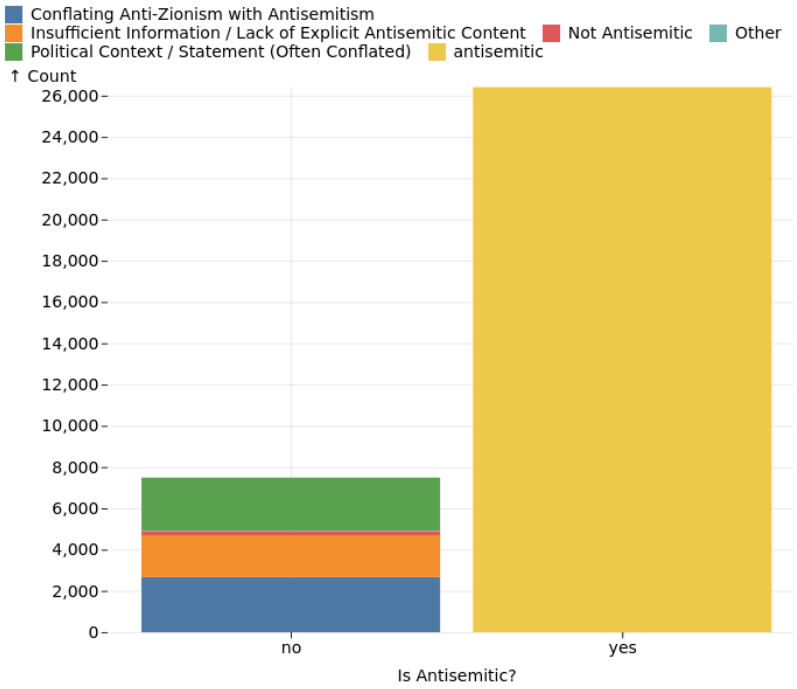
The ADL's antisemitism dataset is quite large, as it contains over 33,000 separate cases. It would be impossible for a single person to evaluate the validity of the ADL's claims of antisemitism. While multiple people could go through the dataset, they would likely interpret the data differently, leading to inconsistent results.

We have decided to use an LLM to classify each example in the ADL's dataset as either antisemitic or not. Furthermore, only data from 2016 to 2024 will be used in this analysis.

To begin our analysis, we computed both lower and upper bound LLM-based estimates for whether each incident in the dataset was antisemitic. For incidents that are not antisemitic, we also asked the LLM to provide a reason why the ADL might think it is (false positive reason).

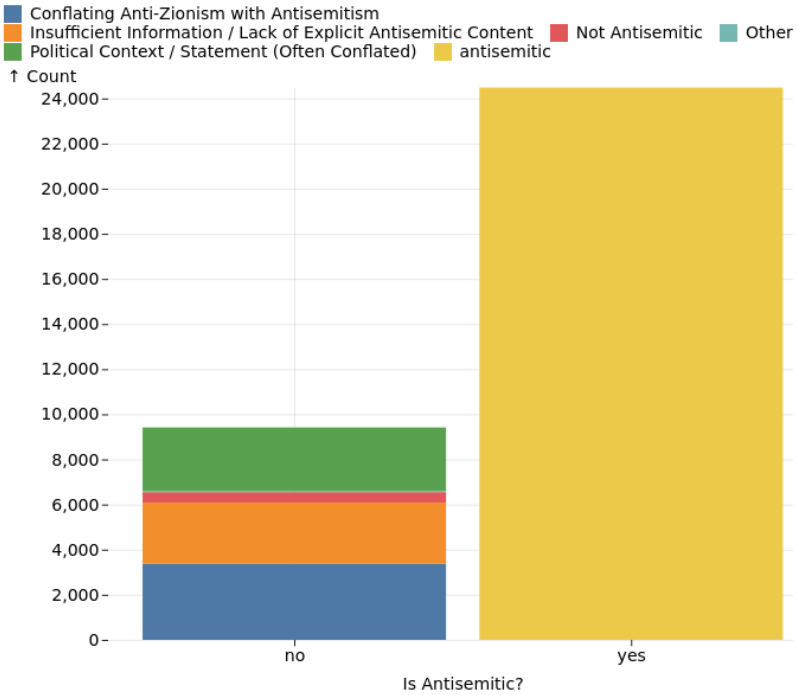
Results – Many recorded incidents are not antisemitic.

In this section, we have computed both lower and upper bound estimates of antisemitism. For the incidents that the models did not determine to be antisemitic, we asked them to provide a reason.



Is Antisemitic Label	False Positive Reason	# Incidents	% All Incidents	% All Incidents Within Label
no	Conflating Anti-Zionism with Antisemitism	2671	7.88%	35.65%
no	Insufficient Information / Lack of Explicit Antisemitic Content	2026	5.98%	27.04%
no	Not Antisemitic	210	0.62%	2.80%
no	Other	19	0.06%	0.25%
no	Political Context / Statement (Often Conflated)	2566	7.57%	34.25%
yes	antisemitic	26412	77.90%	100.00%

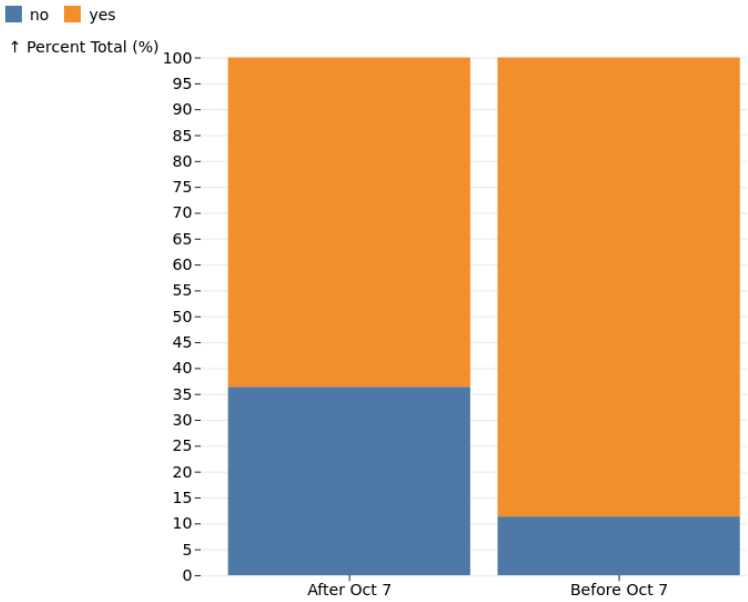
Figure 5. Lower bound estimate for false positive antisemitic incidents across entire ADL dataset



Is Antisemitic Label	False Positive Reason	# Incidents	% All Incidents	% All Incidents Within Label
no	Conflating Anti-Zionism with Antisemitism	3367	9.93%	35.75%
no	Insufficient Information / Lack of Explicit Antisemitic Content	2717	8.01%	28.85%
no	Not Antisemitic	469	1.38%	4.98%
no	Other	41	0.12%	0.44%
no	Political Context / Statement (Often Conflated)	2824	8.33%	29.99%
yes	N/A	24486	72.22%	100.00%

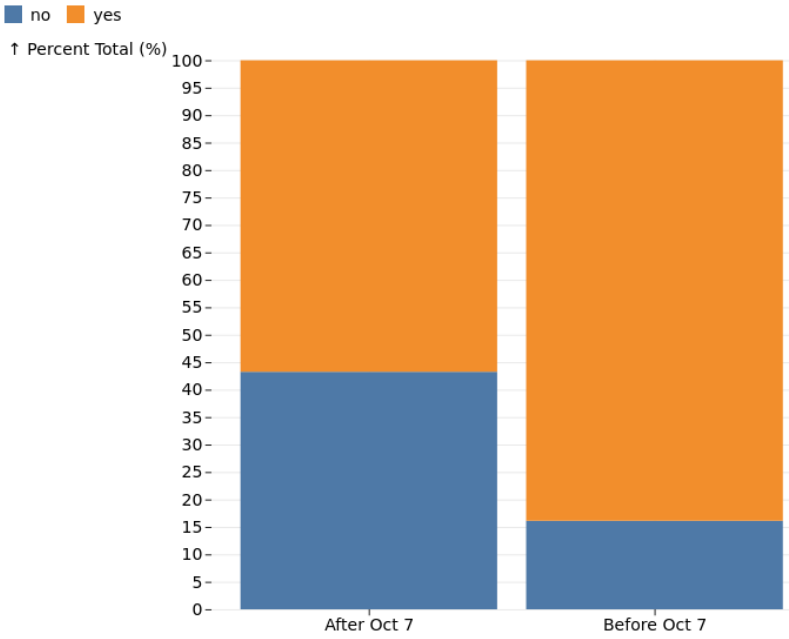
Figure 6. Upper bound estimate for false positive antisemitic incidents across entire ADL datase

We also computed both upper and lower bound estimates of antisemitism, as mentioned above, but for dates before and after October 7, 2023.



Period	Is Antisemitic Label	# Incidents	% All Incidents	% All Incidents Within Period
After Oct 7	no	5298	15.63%	36.39%
After Oct 7	yes	9259	27.31%	63.61%
Before Oct 7	no	2194	6.47%	11.34%
Before Oct 7	yes	17153	50.59%	88.66%

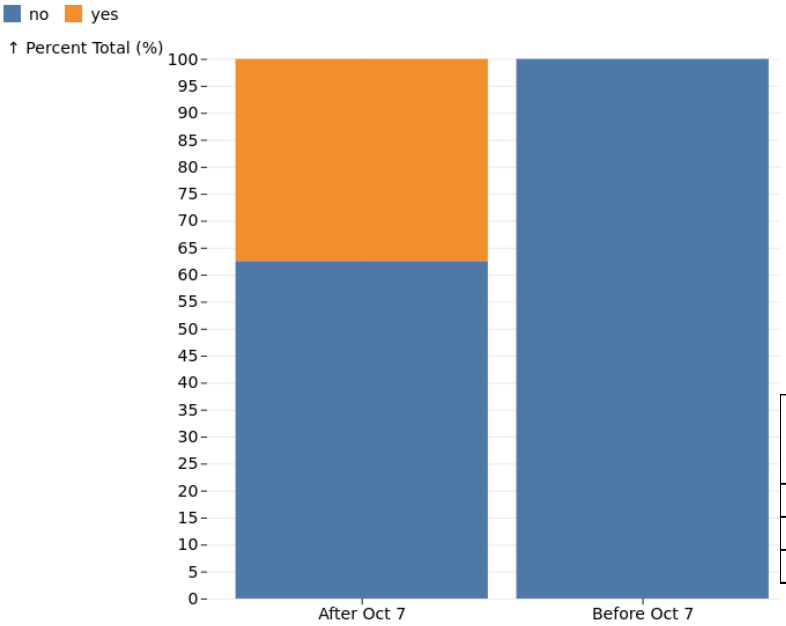
Figure 7. Lower bound estimate for false positive antisemitic incidents before and after October 7 2023



Period	Is Antisemitic Label	# Incidents	% All Incidents	% All Incidents Within Period
After Oct 7	no	6298	18.58%	43.26%
After Oct 7	yes	8259	24.36%	56.74%
Before Oct 7	no	3120	9.20%	16.13%
Before Oct 7	yes	16227	47.86%	83.87%

Figure 8. Upper bound estimate for false positive antisemitic incidents before and after October 7 2023

One way we verified the general correctness of our method was to look at the "is_israel_zionism_related" column available in the ADL's dataset. This column was added by the ADL after October 7, 2023. We can see that after October 7, the number of Israel/Zionism incidents made up **37.45%** of all incidents, which is between our lower and upper bound estimates of **36.39%** and **43.26%**, respectively.



Period	Is Israel Zionism Related?	# Incidents	% All Incidents	% All Incidents Within Period
After Oct 7	no	9105	26.86%	62.55%
After Oct 7	yes	5452	16.08%	37.45%
Before Oct 7	no	19347	57.06%	100.00%

Figure 9. Number of Israel/Zionism related incidents in the ADLs dataset (according to the ADL).

We then verified that our prompt was not just naively labeling examples related to Israel/Zionism as "Not Antisemitic." We did this by computing the joint distribution between the original "is_israel_zionism_related" column available in the ADL's dataset and the lower bound estimate prediction of the antisemitism label. The results show that, after October 7, only **22.83%** of all data points that the ADL labels as related to Israel/Zionism are labeled as "Not Antisemitic" by our lower bound estimate. In other words, our estimate does not naively label examples related to Israel/Zionism as "Not Antisemitic." We attribute this to the many hours spent designing and iterating the prompt.

Period	Is Israel Zionism Related?	Is Antisemitic Label	# Incidents	% All Incidents	% All Incidents Within Period
After Oct 7	False	False	1975	5.83%	13.57%
After Oct 7	False	True	7130	21.03%	48.98%
After Oct 7	True	False	3323	9.80%	22.83%
After Oct 7	True	True	2129	6.28%	14.63%
Before Oct 7	False	False	2194	6.47%	11.34%
Before Oct 7	False	True	17153	50.59%	88.66%

Figure 10. Joint distribution of incidents the ADL labels as related to Israel/Zionism and those our estimates labeled as antisemitic.

We finally computed both upper and lower bound estimates of antisemitism for each month in the data-set from 2016 to 2024.

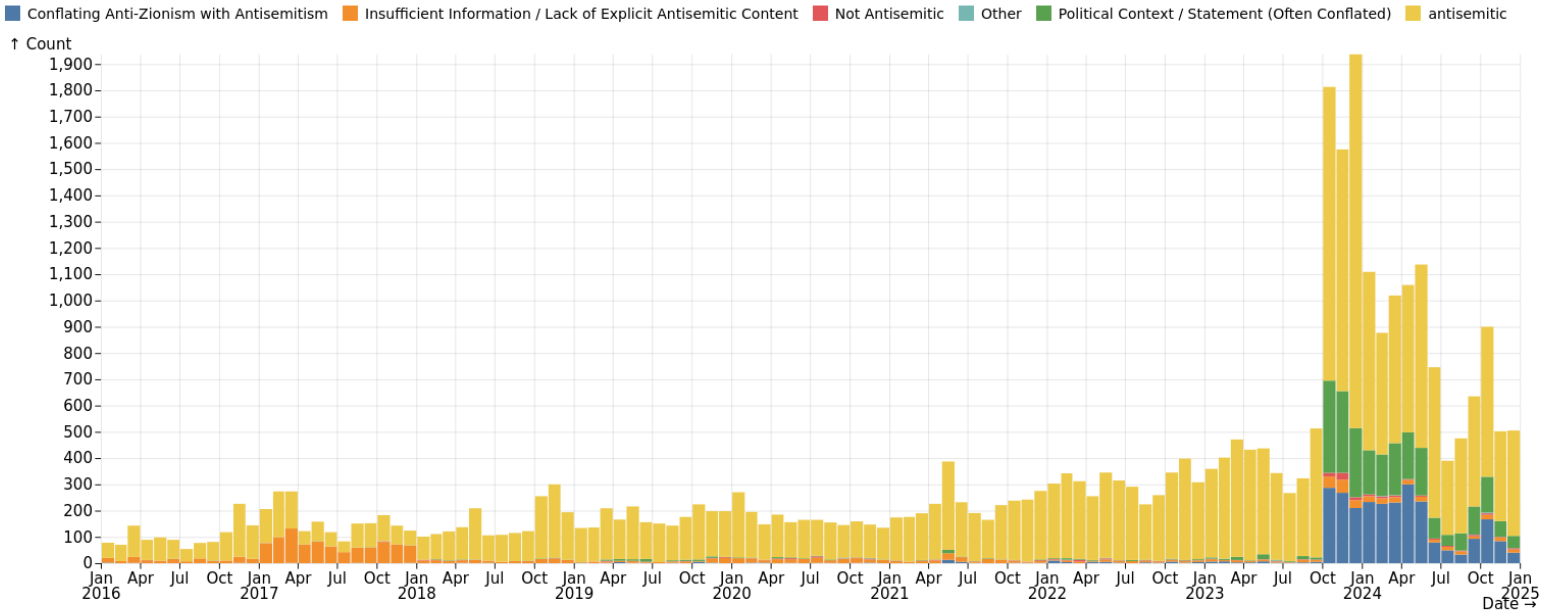


Figure 11. Lower bound estimate for false positive antisemitic incidents from 2016 to 2024

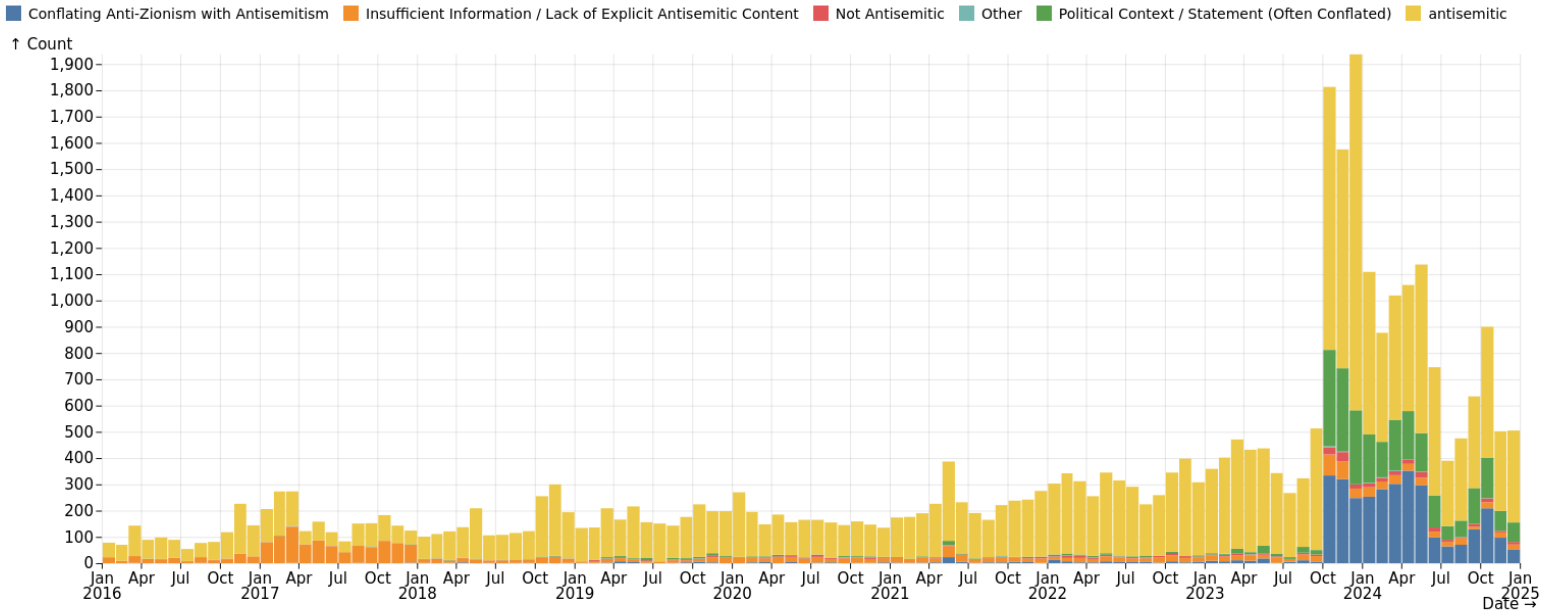


Figure 12. Upper bound estimate for false positive antisemitic incidents from 2016 to 2024

Results – LLMs have a Zionist bias

To understand the current biases of existing LLM models, we decided to run an experiment where we tested multiple examples from the ADL's dataset across various levels of severity/antisemitism (or lack thereof) through all the major LLM providers. We also added our own made-up list of descriptions (marked as 1_a, 1_b, and 1_c in the table below).

We tested the descriptions across two different conditions. The first condition, without guidelines, is where we asked the LLM whether it thought a description was antisemitic or not, without providing explicit instructions, guidelines, or examples. The second condition, with guidelines, is the opposite of the first. In this condition, we provided explicit instructions on things to look for (e.g., clear antisemitic examples) and avoid (e.g., avoiding conflating antisemitism with anti-Zionism). A copy of the prompt used for both conditions can be found in the methods section.

Note example 1_b. We noticed that when example 1_c was omitted from the batch, all LLMs labeled 1_b as "antisemitic." When 1_c was added, all LLMs, with the exception of Deepseek, labeled it as "not antisemitic." We refer to this as "context contamination" or "context influencing," where the LLM prediction for the current item in the batch can be influenced by the presence of other items in the batch. We discuss this phenomenon further in the methods section below, titled "Handling Inconsistent Cases."

		Without Guidelines				With Guidelines			
Incident ID	**Incident Description**	**GPT 4o**	**Llama 4 70B**	**Deekseek V3**	**Gemini 2.5 F**	**GPT 4o**	**Llama 4 70B**	**Deekseek V3**	**Gemini 2.5 F**
65386	At an anti-Israel rally organized by Students for Socialism, Students for Justice in Palestine, the Party for Socialism and Liberation and Stop Cop City at Emory University, protesters chanted, "Zionist Zionist go back home, Palestine is our home." A speaker also referred to the University's president as a "Zionist puppet."	Yes	Yes	Yes	Yes	No	No	No	Yes
62356	An individual received antisemitic text messages.	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
68564	Graffiti was found on a building at Vassar College that read: "Intifada," a slogan commonly used as a call for violence against Israel and the Jewish community. The vandalism also included inverted red triangles, a symbol that first appeared in Hamas propaganda videos showing Israeli military targets.	Yes	Yes	Yes	Yes	Yes	No	No	No
61647	A synagogue was targeted with an antisemitic bomb threat.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
62259	"From the river to the sea Palestine will be free!" was written in chalk on the campus of the University of Kansas.	Yes	Yes	Yes	Yes	No	No	No	No
57436	An unknown student posted a message about their teacher referring to them as a "raging Zionist" and writing: "Fuck him."	Yes	Yes	Yes	Yes	No	No	No	No
28354	A visibly Jewish man was assaulted by a man on the street who right before the assault said, "Keep on walking you Jewish motherfucker. I should beat your Jewish motherfucking ass right now."	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
68996	At an anti-Israel rally organized by the Arab Resource & Organizing Center, the Do No Harm Coalition and San Francisco Rising Action, protesters chanted, "From the river to the sea, Palestine will be free," a slogan commonly used to call for an end to the Jewish state.	Yes	Yes	Yes	Yes	No	No	No	No
68343	Patriot Front, a white supremacist group, posted stickers that read: "One nation, against invasion," "No Zionists in government, we serve one nation," "Strong families, strong nations" and "Defend American labor."	Yes	Yes	Yes	Yes	No	Yes	No	Yes
5924	A seller on Amazon received a harassing message from a buyer, in which the buyer called the seller a "kike."	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
1_a	A group known as the AAA protested screaming, “Israel Israel stop the genocide”.	No*	No*	Yes*	No*	No	No	No	No
1_b	A known antisemitic group the BBB protested screaming, “Israel Israel stop the genocide”	Yes	Yes	Yes	Yes	No	No	No	No
1_c	A group known as CCC protested chanting, "Boycott Israel!”	No	No	No	No	No	No	No	No

Figure 13. Antisemitism classification of various LLMs for different incident descriptions. Results available for LLM inference done without and with additional guidelines added to the prompt.

Methodology

Data availability

All data used and generated in this investigation is available for download.

adl_2016_to_2024_antisemitism_data_labeled (TSV) (XLSX)

- Joined table between ADLs data and LLM classification labels
- Contains data from 2016 to 2024. Some rows omitted where LLMs failed to provide proper structured output

prompt_plain.md

- Contains the prompt used for assessing biases across various LLM providers

prompt_guidelines.md

- Contains the final prompt used in the investigation

Data collection and observations

Data on antisemitic incidents in the United States were obtained from the ADL's website. Antisemitic incident data is only available in abundance from 2016 to 2024. Therefore, the rest of the investigation will only use data from that period.

We verified the number of incidents from the dataset with the official reporting of the ADL. The yearly number of incidents is almost an exact match.

Year	# Rows in ADL's dataset	# Incidents reported by ADL
2023	8872	8873
2022	3698	3698
2021	2717	2717
2020	2025	2026
2019	2107	2107
2018	1879	1879
2017	1986	1986
2016	1267	1267

Figure M1. Comparing the number of rows available in ADL's dataset and the number of incidents reported by the ADL

We inspected their data and made the following observations:

- There is a column available called israel_zionism_related. A value of 1 in this column suggests that the incident being described is in some way related to either Israel or Zionism.
- There is a column available called image, which contains a link leading to archive.org with a picture of the alleged incident. In the dataset provided, 98.5% of the data do not have a link or image available. Of the rows that contain a link to the image, only 320 of the 496 (64.5%) contain a unique link. In other words, there are 176 rows with duplicate links despite the incident being recorded as different. Upon further inspection, we noticed that this occurs in cases where similar incidents happen in multiple areas across multiple points in time. For example, a white supremacist group distributes racist posters. We also found that the majority of links no longer lead back to a valid URL, with the following message being returned when trying to access them: “The Wayback Machine has not archived that URL.”
- Many of the alleged antisemitic incidents do not contain a clear description of what exactly happened. For example, consider case id=63796: “An individual yelled antisemitic language in a public area.” There is no additional context provided about this. A large subset of the ADL’s data is like this. It’s not that they are unable to provide specific descriptions, as they have done so for other cases, such as id=11719: “Fliers that read 'Zionism is Neo-Nazism' were found posted throughout the UC Berkeley campus.”

Using LLMs as a classifier of antisemitism

This is where most of the effort of this investigation went. LLMs, without guidance, do not accurately classify whether something is antisemitic or not. This is evident in the section above, where we showcased that the LLMs exhibit a Zionist bias and label anything with anti-Israel sentiment as antisemitic. However, with guidance, LLMs become much more capable of accurately classifying text.

In the generative AI world, prompts can be used to provide additional context or instructions to the LLM. In our case, we spent hours iterating on our prompt more than 30 times to ensure that the results reflect the goal of this investigation, which is to have an AI classifier that is as objective as possible (with respect to the guidelines) and understands the nuances of biases present in the data.

Cases that conflate anti-Zionism with antisemitism

The most common type of false-positive case, as seen in the investigation above, is the conflation of anti-Zionism with antisemitism. There are countless examples of this in the dataset, such as example 1 (id=713): “Individual received an anti-Semitic email claiming Israel was carrying out a genocide against refugees.”

With no clear evaluation guidelines provided, almost all LLMs tested (Deepseek R1, GPT-04 mini, Gemini 2.5 Pro) classify anti-Zionist statements as antisemitic. To address this, we have added section

6 to our evaluation guidelines, “Anti-Zionism vs. Antisemitism,” where we explain what the difference is (see Final Prompt Section below).

Cases with leading descriptions

One common type of case that leads to false positives is leading descriptions. These are descriptions where the authors from the ADL label something as antisemitic to demonstrate that it is. It is like labeling a random object as "dangerous" and then pointing to that label as proof of its threat, without any real evidence to support the claim. Below is a list of examples:

- Example 1 (id=3525): “Conspiracy-filled anti-Semitic and anti-Israel flyers discovered in a parking lot.”
- Example 2 (id=3650): “Anti-Semitic Incident: Vandalism on Walk for Israel website.”
- Example 3 (id=65114): “At an anti-Israel rally organized by the Party for Socialism and Liberation at Georgia Tech, protesters displayed signs with messages that included: 'Resistance against occupation is a human right' and 'From the river to the sea,' slogans commonly used to support Hamas's antisemitic October 7, 2023, terrorist attack and as a call for an end to the Jewish state.”

Example 1 mentions that the flyer is “anti-Israel”. Is the flyer antisemitic because it is anti-Israel, or because it contains both antisemitic and anti-Israel content?

Example 2 mentions that the vandalism of a website is an “Anti-Semitic Incident”. Did the vandalism contain antisemitic messaging, or did the ADL consider it antisemitic because the vandalism was related to Israel?

Example 3 mentions the phrases "Resistance against occupation is a human right" and "From the river to the sea" are “slogans commonly used to support Hamas's antisemitic October 7, 2023, terrorist attack”. This is probably the most egregious example, as these phrases were commonly used even prior to October 7, 2023. Furthermore, labeling attacks against Israel as inherently antisemitic is another example of this.

All three of these examples contain leading descriptions by the ADL. Many times, the LLMs would label these incidents as antisemitic simply because the ADL suggested that they were. In other words, the justification of antisemitism was in the ADL's labeling of it.

To combat this, we added the following instructions to the prompt: “Ignore the inflammatory nature of language unless the described content specifically meets criteria in Section 2”. In other words, the incident itself has to meet the criteria for antisemitism, as opposed to the ADL’s comments about the incident.

Cases that lack information or are ambiguous

Another common type of case that leads to false positives is ambiguity. These are descriptions where minimal evidence is provided about the incident. Below is a list of examples:

- Example 1 (id=57878): “An individual made an antisemitic comment to their neighbor.”
- Example 2 (id=69286): “Anti-Israel protesters organized by the antisemitic and anti-Israel group Jewish Voice for Peace disrupted an event hosted inside a synagogue.”
- Example 3 (id=17159): “A Jewish man was physically assaulted on the street in an unprovoked attack.”

None of the examples explicitly state what was said, and the intent behind the actions is unclear. Depending on the type of ambiguity presented in these examples, our prompt suggests handling each case differently.

In Example 1, our prompt indicates that this is an antisemitic case according to the guidelines in Section 5, “Handling Ambiguity.” For straightforward cases like this, where antisemitism is identified as the issue without any conflation (such as anti-Zionism), we give the benefit of the doubt and label it as antisemitism.

In Example 2, the LLMs would not classify this as antisemitic due to the conflation with anti-Israel protests. However, our lower bound estimate considers any incident containing the following words as antisemitic as a precaution: synagogue, kippah, rabbi. In reality, this case would require additional context to determine if it aligns with the type of protest that “Jewish Voice for Peace” has conducted to disrupt events related to land sales in the West Bank. Given the complexity of this case, and recognizing the limitations of LLMs, we are willing to accept the risk of false positive penalties rather than allowing for false negatives.

In Example 3, we do not consider the individual's identity to be relevant, as no motive for the attack has been provided. Therefore, this case would be labeled as not antisemitic.

Cases where LLMs are inconsistent

As mentioned in the investigation above, there are cases where the LLM prediction for the current item in the batch can be influenced by the presence of other items in the batch. We refer to this as “context contamination” or “context influencing.”

When we first investigated this, we thought it was a case of LLM stochasticity and randomness. We tried setting the temperature parameter to 0, but that did not help. After changing the prompt more than a dozen times to no avail, we realized that this was not a system prompt issue, but an issue with batch inference with LLMs. Once we fed both examples individually into the LLM (batch size of 1), the LLM predicted “antisemitic” for both.

Upon further investigation, we realized we were probably running into context contamination, where the presence of other examples in the batch directly influenced the results of this one example. We looked into this topic more and found that there were other people experiencing the same issue. We learned that LLMs have an inherent instability, and this is the reality of running batch inference with LLMs [1, 2].

The outcome of this behavior is similar to the "contrast effect" in humans. If I am presented with 100 examples that are much worse than example 1, I am less likely to think that example 1 is problematic. Similarly, if I am presented with 100 examples that are less offensive than example 1, I may perceive example 1 as more severe than it actually is.

Below is a list of examples for which we encountered this phenomenon:

- Example 1 (id=68918): "Patriot Front, a white supremacist group, posted stickers that read: 'American spirit, European blood,' 'No Zionists in government, we serve one nation,' 'Reclaim America,' and 'Life, liberty, victory.'"
- Example 2 (id=68911): "Patriot Front, a white supremacist group, posted stickers that read: 'No Zionists in government, we serve one nation,' 'Reject poison,' and 'One nation, against invasion.'"

Despite being essentially the same kind of incident, example 1 got flagged as antisemitic while example 2 did not. When asked for the reason why the LLM thought this way, the LLM replies:

- Example 1 (id=68918): "Propaganda from a white supremacist group explicitly stating 'No Zionists in government, we serve one nation.' While 'Zionist' can be political, from a white supremacist group, this phrase often functions as a stand-in for 'Jews' exercising undue influence, falling under Guideline 2B via association with a group defined by anti-Jewish ideology and the 'control' trope. Confidence is medium as the term 'Zionist' is used, not 'Jew.'"
- Example 2 (id=68911): "Description details nationalistic/political slogans ('No Zionists in government,' 'One nation, against invasion'). Objective details do not contain explicit antisemitic content as defined in Section 2. Disregard of source group's affiliation unless content directly meets criteria (Guideline 3)."

The next section will explain how we handled inconsistent cases.

Handling inconsistent cases

We explored various strategies to address the inconsistencies in large language model (LLM) outputs that arise during batch inference. While the most precise solution would be to set the batch size to one, this approach would be quite costly in terms of computational resources.

Our observations revealed that inconsistent cases often emerged in instances that teetered on the edge of being antisemitic or when the context was not quite clear from the case descriptions. Although we

found that incorporating clear examples into our system prompt alleviated some of these issues, the sheer volume of potential edge cases made it impractical to hard-code every scenario. Moreover, we do not position ourselves as experts in these nuanced matters concerning antisemitism, and thus prefer not to force our opinions and thoughts as the correct ones.

After experimenting with several approaches, we considered three methods to handle these inconsistent cases:

1. **Labeling Inconsistent Cases as "Inconclusive"**: To determine whether a label is inconsistent, we would need to run LLM inference multiple times. Cases that yield differing labels across two runs would be classified as "inconclusive." Each run would involve shuffling the dataset with a different random seed before inference begins.
2. **Upper Bound Estimate**: In this method, we conducted three separate sets of LLM batch inferences. Similar to the first method, the dataset would be shuffled with a different random seed for each run. The final label would be determined by the most frequently assigned label across all three runs. In other words, a case would be classified as antisemitic if at least two out of the three LLMs identified it as such.
3. **Lower Bound Estimate**: This approach mirrors the second method, but instead of relying on the majority label, the final designation would be the first positive label encountered in the multiple inference runs. Additionally, we compiled a list of keywords specific to Judaism, such as "Synagogue," "Kippah," and "Rabbi." If any of these terms appeared in a case description, we would override a negative label with a positive one. Thus, a case would be deemed antisemitic if at least one of the three LLMs labeled it as such or if any of the designated keywords were present.

In the end, we decided to compute and present both upper bound and lower bound estimates in this investigation.

Handling inconsistent output structures

Despite using a structured output format and specifying the desired output schema in the prompt, there are still times when the LLM produces a broken schema or suggests categories beyond the ones specified in the prompt.

To mitigate this, we grouped false positive reason categories based on similarity. This was done with the help of an LLM. The false positive reason categories were then grouped using these new grouping definitions. To reproduce the groupings presented in this investigation, we have provided a SQL clause that can be used to transform the "false_positive_reason" column into its final grouped form.

```
CASE
    -- 1. Conflating Anti-Zionism with Antisemitism
    WHEN false_positive_reason = 'Conflating Anti-Zionism with Antisemitism' THEN 'Conflating Anti-Zionism with Antisemitism'


    -- 2. Political Context / Statement (Often Conflated)
    WHEN false_positive_reason IN (
      'Political Statement',
      'Conflating Political Group with Antisemitism',
      'Political Group',
      'Conflating Political Statement with Antisemitism',
      'Conflating Political Criticism with Antisemitism',
      'Political Group Statement',
      'Political Group/Statement',
      'Political Group with Antisemitism',
      'Political Group Focus',
      'Political Group with Nationalism',
      'Conspiracy Theory/Political Statement',
      'Political Group Mention',
      'Political Group Conflated',
      'Political Group Membership Not Sufficient',
      'Political Group Affiliation',
      'Conflating Support for a Political/Militant Group with Antisemitism',
      'Conflating Support for Terror Group with Antisemitism',
      'Political Group with Political Statement',
      'Political Group, Not Explicitly Antisemitic Content',
      'Political/Military Symbol',
      'Holocaust Distortion/Equivalency (Political)',
      'Conflating Holocaust Comparison with Antisemitism',
      'Political Group/Other',
      'Political Group Association',
      'Conflating Political Criticism with Harassment',
```

```
        'Political Group Not Explicitly Antisemitic in Description',
        'Political Group, Lack of Information', -- Grouping political lack of info here
        'Political Group Actions',
        'Discrimination based on Perceived Nationality/Association' -- Often political context
    ) THEN 'Political Context / Statement (Often Conflated)'

    -- 3. Insufficient Information / Lack of Explicit Antisemitic Content
    WHEN false_positive_reason IN (
        'Lack of Information',
        'Confidential/Lack of Information',
        'Confidential',
        'Lack of Explicit Antisemitic Content',
        'Lack of Antisemitic Content',
        'Lack of Antisemitic Context',
        'Threat Lacks Antisemitic Content',
        'Lack of Context/Motive',
        'Lacks Explicit Antisemitic Content'
    ) THEN 'Insufficient Information / Lack of Explicit Antisemitic Content'

    -- 4. Not Antisemitic
    WHEN false_positive_reason IN (
        'Not Antisemitic',
        'Not Antisemitic (Political message)', -- Explicitly "Not Antisemitic" takes precedence
        'Not Antisemitic (Lack of direct link)'
    ) THEN 'Not Antisemitic'

    -- 5. null (when not false positive)
    WHEN false_positive_reason IS null THEN null

    -- 6. Other (Catch-all for remaining specific low-count or truly 'Other' categories)
    ELSE 'Other'
END AS simplified_false_positive_reason
```

Final LLM Prompt

Below is the final prompt used for the results of our investigation.

Task: Classify Antisemitic Incidents from Descriptions

Objective: Process a list of incident reports. For each report, analyze the provided `description` based on the **Antisemitism Evaluation Guidelines** below. Prioritize objective details when available. Handle vague descriptions according to Guideline 5, giving specific treatment to simple claims of 'antisemitism' versus those potentially conflated with political criticism. Determine the specified output fields for each incident independently.

Output Format: Provide the results as a JSON list, where each object in the list corresponds to one incident report and contains the following fields:

```
[
  {
    "description_id": "...",
    "is_antisemitic": "yes" | "no",
    "label_confidence": 1 | 2 | 3,
    "description_specificity": 1 | 2 | 3,
    "event_category": "...",
    "action_category": "...",
    "false_positive_reason": "..." | null, // String value or null
    "is_antisemitic_reason": "..."
  },
  {
    // ... next incident
  }
]
```

Field Definitions:

- **`description_id`**: Unique identifier (from input brackets).
- **`is_antisemitic` (yes/no)**: Based on objective `description` details vs. Guidelines. See Guideline 5 for handling vague descriptions and submitter labels.
- **`label_confidence` (1/2/3)**: Confidence in `is_antisemitic` classification. (See Confidence Definitions below). Note: Level 1 applies to both "no" due to lack of info *and* "yes" based on the 'benefit of the doubt' rule for simple vague claims.
- **`description_specificity` (1/2/3)**: How specific and detailed is the objective information *about the incident's content* within the `description`? (See Specificity Definitions below).
- **`event_category`**: Setting/context (e.g., `Educational Institution`, `Public Area`, `Online Communication`). Choose the most appropriate.
- **`action_category`**: Type of act (e.g., `Vandalism`, `Verbal Harassment`, `Display of Symbols`, `Protest/Demonstration`). Choose the most appropriate.

- **`false_positive_reason`**: If `is_antisemitic` is "no", categorize why (e.g., `Conflating Anti-Zionism with Antisemitism`, `Conflating Political Group with Antisemitism`, `Not Antisemitic`, `Political Statement`, `Other`). **MUST be a string if `is_antisemitic` is "no", MUST be `null` if `is_antisemitic` is "yes".**
- **`is_antisemitic_reason`**: Brief explanation linking description details (or lack thereof) to the Guidelines for the "yes" or "no" classification. Justify based on objective details or the specific handling of vagueness per Guideline 5.

Examples:

Example 1 (Clear Antisemitism, High Specificity): *Input Description:* `[ID-001] Saw graffiti on the synagogue wall this morning. It included a large swastika and the words 'Jews Out'.` *Expected Output Snippet:*

```
{
  "description_id": "ID-001",
  "is_antisemitic": "yes",
  "label_confidence": 3,
  "description_specificity": 3, // Specific symbol and direct quote provided
  "event_category": "Religious Institution",
  "action_category": "Vandalism",
  "false_positive_reason": null, // Is_antisemitic is yes
  "is_antisemitic_reason": "Explicit antisemitic symbol (swastika) and slogan
('Jews Out') described (Guideline 2A/2B)."
}
```

Example 2 (Political Speech, Not Antisemitic, High Specificity): *Input Description:* `[ID-002] A protest march criticizing Israeli government actions in Gaza. Signs said 'Free Palestine' and 'Stop the Occupation'. One person briefly chanted 'From the river to the sea'. I felt it was antisemitic.` *Expected Output Snippet:*

```
{
  "description_id": "ID-002",
  "is_antisemitic": "no",
  "label_confidence": 3,
  "description_specificity": 3, // Specific sign text and chant content provided
  "event_category": "Protest/Demonstration",
  "action_category": "Protest/Demonstration",
  "false_positive_reason": "Conflating Anti-Zionism with Antisemitism", // is_anti-
semitic is no
  "is_antisemitic_reason": "Description details political criticism of Israel. No
objective details meet antisemitism criteria (Guideline 3 & 6)."
}
```

Example 3 (Insufficient Information, Low Specificity - General Hate): *Input Description:* `[ID-003] My neighbor yelled something hateful at me yesterday. He knows I'm Jewish.` *Expected Output Snippet:*

```
{
  "description_id": "ID-003",
  "is_antisemitic": "no", // "Hateful" is vague, not explicitly "antisemitic" label
  "label_confidence": 1,
  "description_specificity": 1, // Vague description ('something hateful'), lacks
content details
  "event_category": "Residential Area",
  "action_category": "Verbal Harassment", // (Alleged)
  "false_positive_reason": "Lack of Information", // is_antisemitic is no
  "is_antisemitic_reason": "Description states 'something hateful' but lacks ob-
jective details (specific words, slurs, tropes) or direct 'antisemitic' label. Can-
not verify antisemitism per guidelines (Guideline 5)."
}
```

Example 4 (Inflammatory Political Slogan, Not Antisemitic per Guidelines, High Specificity): *Input Description:* `[ID-004] At an anti-Israel encampment, protesters displayed signs with messages that included: 'Zionism is fascism.'` *Expected Output Snippet:*

```
{
  "description_id": "ID-004",
  "is_antisemitic": "no",
  "label_confidence": 3,
  "description_specificity": 3, // Specific sign text provided
  "event_category": "Educational Institution",
  "action_category": "Display of Signs",
  "false_positive_reason": "Conflating Anti-Zionism with Antisemitism", // is_anti-
semitic is no
  "is_antisemitic_reason": "Description details inflammatory political criticism of
Zionism. Does not meet specific antisemitism criteria per guidelines (Guideline
6)."
}
```

Example 5 (Controversial Chant, Not Antisemitic per Guidelines, High Specificity): *Input Description:* `[ID-005] At an anti-Israel rally, protesters chanted, 'There is only one solution; Intifada, revolution' and 'Intifada, Intifada; Long live the Intifada'.` *Expected Output Snippet:*

```
{
  "description_id": "ID-005",
  "is_antisemitic": "no",
  "label_confidence": 2, // Medium confidence due to borderline nature of applying
guidelines
  "description_specificity": 3, // Specific chant content provided
  "event_category": "Protest/Demonstration",
  "action_category": "Chanting",
  "false_positive_reason": "Conflating Anti-Zionism with Antisemitism", // is_anti-
semitic is no
```

```
  "is_antisemitic_reason": "Description details controversial political chants
('Intifada'). Objective details do not explicitly include calls for violence
against Jews as a group or other antisemitic content required by Guideline 6 or
Section 2."
}
```

Example 6 (Medium Specificity, Antisemitic): *Input Description:* `[ID-006] Someone drew a symbol on the door that looked like a swastika, but it was messy. They also yelled something about Jews and money.` *Expected Output Snippet:*

```
{
  "description_id": "ID-006",
  "is_antisemitic": "yes", // Meets criteria via trope/symbol despite imprecision
  "label_confidence": 2, // Medium confidence due to imprecision ("looked like",
"something about")
  "description_specificity": 2, // Some detail, but lacks precision
  "event_category": "Residential Area", // Or Public Area depending on 'door' con-
text
  "action_category": "Vandalism", // And Verbal Harassment
  "false_positive_reason": null, // is_antisemitic is yes
  "is_antisemitic_reason": "Description mentions symbol resembling swastika and
paraphrased classic trope (Jews/money), meeting Guideline 2B, though details lack
precision."
}
```

Example 7 (Vague Claim - Simple "Antisemitic" Label): *Input Description:* `[ID-007] Received online feedback containing an antisemitic comment.` *Expected Output Snippet:*

```
{
  "description_id": "ID-007",
  "is_antisemitic": "yes", // Applying 'benefit of the doubt' for simple, unquali-
fied 'antisemitic' label (Guideline 5 exception)
  "label_confidence": 1, // Low confidence due to lack of specific details
  "description_specificity": 1, // Vague description ('antisemitic comment')
  "event_category": "Online Communication",
  "action_category": "Communication",
  "false_positive_reason": null, // is_antisemitic is yes
  "is_antisemitic_reason": "Description provides no specific content but uses
simple 'antisemitic' label without other context. Applying Guideline 5 exception:
classifying as 'yes' with low confidence due to lack of detail."
}
```

Example 8 (Vague Claim - Simple "Antisemitic" Label, Institution Context): *Input Description:* `[ID-008] A Jewish community center reported receiving an antisemitic voicemail message.` *Expected Output Snippet:*

```
{
  "description_id": "ID-008",
  "is_antisemitic": "yes", // Applying 'benefit of the doubt' for simple, unquali-
fied 'antisemitic' label (Guideline 5 exception)
  "label_confidence": 1, // Low confidence due to lack of specific details
  "description_specificity": 1, // Vague description ('antisemitic voicemail')
```

```
  "event_category": "Religious Institution", // Or Community Center
  "action_category": "Communication", // Or Verbal Harassment
  "false_positive_reason": null, // is_antisemitic is yes
  "is_antisemitic_reason": "Description provides no specific content but uses
simple 'antisemitic' label without other context. Applying Guideline 5 exception:
classifying as 'yes' with low confidence due to lack of detail."
}
```

Example 9 (Vague Claim - Explicit Conflation Suspected): *Input Description:* `[ID-009] Found a flyer posted on campus described as 'antisemitic and anti-Israel'.`
Expected Output Snippet:

```
{
  "description_id": "ID-009",
  "is_antisemitic": "no", // Benefit of doubt NOT applied due to potential confla-
tion signaled by 'anti-Israel' pairing. Falls back to default lack of info rule.
  "label_confidence": 1, // Low confidence, classification based on lack of verifi-
able detail
  "description_specificity": 1, // Vague description, lacks content details
  "event_category": "Educational Institution",
  "action_category": "Propaganda Distribution",
  "false_positive_reason": "Conflating Anti-Zionism with Antisemitism", // is_anti-
semitic is no
  "is_antisemitic_reason": "Description lacks specific content. The phrase 'anti-
semitic and anti-Israel' suggests potential submitter conflation, preventing the
Guideline 5 exception. Defaulting to 'no' due to lack of verifiable objective de-
tails (Guideline 5)."
}
```

```
**Example 10 (Threat Targeting Jewish Institution, Medium Specificity/
Confidence):**
*Input Description:* `[ID-010] A Jewish community center received a bomb threat via
email.`
*Expected Output Snippet:*
```json
{
  "description_id": "ID-010",
  "is_antisemitic": "yes", // Targeting a Jewish institution with a bomb threat
meets Guideline 2A
  "label_confidence": 2, // Medium confidence: Act (threat) is clear, but specific
content/language of the threat email is unknown, preventing high confidence.
  "description_specificity": 2, // Medium specificity: Action (bomb threat via
email) is specific, but the description lacks details on the *content* of the email
itself.
  "event_category": "Religious Institution", // Or Community Center
  "action_category": "Threat", // Or Communication
  "false_positive_reason": null, // is_antisemitic is yes
  "is_antisemitic_reason": "Description details a specific threat (bomb threat) ex-
plicitly targeting a Jewish institution. This qualifies under Guideline 2A (target-
ing) even without the specific wording of the threat. Lack of email content details
results in medium confidence/specificity."
}
```
```

Example 11 (Assault on Identifiable Jew at Jewish Location): *Input Description:* [ID-011] An Orthodox rabbi was assaulted with a stick by an individual in front of a synagogue. *Expected Output Snippet:*

```
{
  "description_id": "ID-011",
  "is_antisemitic": "yes", // Assault targeting a visibly identifiable Jew (Orthodox rabbi) at a Jewish site (synagogue) meets Guideline 2A.
  "label_confidence": 2, // Medium confidence: While context strongly suggests antisemitic motive per Guideline 2A, the description lacks explicit confirmation of motive or accompanying antisemitic statements.
  "description_specificity": 2, // Medium specificity: Describes the victim, location, and action clearly, but lacks details on motive or specific words uttered during the assault.
  "event_category": "Religious Institution", // Or Public Area
  "action_category": "Physical Assault",
  "false_positive_reason": null, // is_antisemitic is yes
  "is_antisemitic_reason": "Description details assault on a visibly identifiable Jewish individual (Orthodox rabbi) outside a Jewish institution (synagogue). This context qualifies under Guideline 2A (targeting) even without explicitly stated motive or slurs. Confidence is medium due to lack of explicitly stated motive."
}
```

Example 12 (Vandalism Targeting Synagogue with Offensive Materials): *Input Description:* [ID-012] Synagogue vandalized with pig intestines, wax smeared on the building's window, and broken glass was found at the entrance. *Expected Output Snippet:*

```
{
  "description_id": "ID-012",
  "is_antisemitic": "yes", // Vandalism using materials deeply offensive in a Jewish context (pig intestines) directed at a synagogue strongly implies hostility targeting Jews (Guideline 2A).
  "label_confidence": 3, // High confidence: The specific, offensive nature of the vandalism targeting a synagogue provides clear evidence of antisemitic intent under Guideline 2A.
  "description_specificity": 3, // High specificity: The specific acts of vandalism and materials used are clearly described.
  "event_category": "Religious Institution",
  "action_category": "Vandalism",
  "false_positive_reason": null, // is_antisemitic is yes
  "is_antisemitic_reason": "Description details vandalism targeting a synagogue using materials (pig intestines) highly offensive in a Jewish context. This strongly indicates hostility based on religious identity, meeting Guideline 2A. Specific details provided lead to high confidence."
}
```

Example 13 (Collective Blame Directed at Jewish Entity): *Input Description:* [ID-013] A Jewish non-profit received an email that read: "SHAME ON YOU FREE PALESTINE YOU HAVE BLOOD ON YOUR HANDS." *Expected Output Snippet:*

```
{
  "description_id": "ID-013",
```

```
  "is_antisemitic": "yes", // Directing accusations related to Israel/Palestine
('blood on your hands') at an entity identified *only* as 'Jewish' constitutes
holding Jews collectively responsible (Guideline 3/6).
  "label_confidence": 2, // Medium confidence: Fits the guideline for collective
responsibility based on targeting context, but lacks explicit antisemitic slurs/
tropes beyond the implied collective blame.
  "description_specificity": 3, // High specificity: The exact text of the email
and the target type are provided.
  "event_category": "Online Communication", // Or Commercial Establishment
  "action_category": "Harassment", // Or Communication
  "false_positive_reason": null, // is_antisemitic is yes
  "is_antisemitic_reason": "Targets an entity identified solely as 'Jewish non-
profit' with accusations ('blood on your hands') related to the Israeli-Palestinian
conflict ('Free Palestine'). This holds a Jewish entity collectively responsible
for actions perceived to be related to Israel, meeting Guideline 3/6 criteria for
antisemitism. Confidence is medium as it relies on targeting context rather than
explicit slurs."
}
```

Description Specificity Level Definitions

- **1 (Low Specificity):**
    - **Reasoning:** The `description` provides minimal or vague objective details about the actual content of the incident (what was said, written, displayed, or done). It lacks concrete information needed for clear evaluation against the Guidelines.
    - **Characteristics:** Uses generic terms like "offensive symbol," "hateful comment," "antisemitic statement," without specifying the symbol, comment, or statement. Relies heavily on the submitter's assertion or interpretation of content rather than describing the content itself.
    - **Example Scenario:** "An individual made antisemitic remarks." (What remarks?) "There was offensive graffiti." (What did it depict or say?)
- **2 (Medium Specificity):**
    - **Reasoning:** The `description` provides some concrete objective details about the incident's content, but these details may be partially incomplete, paraphrased, slightly ambiguous, or lack full precision.
    - **Characteristics:** Mentions specific elements but with uncertainty (e.g., "symbol *looked like* a swastika," "graffiti *might have said* X"). Paraphrases speech or text instead of providing direct quotes (e.g., "yelled *something about* Jews and money"). Describes an action but lacks full context needed to be completely unambiguous based on description alone.
    - **Example Scenario:** "The sign mentioned Zionists and blood." (Paraphrased, lacks exact phrasing). "He drew a symbol similar to those used by Nazis." (Vague comparison).
- **3 (High Specificity):**

- **Reasoning:** The `description` provides clear, explicit, and concrete objective details about the incident's content. It offers sufficient information to evaluate directly against the Guidelines with minimal ambiguity *about what occurred or was communicated*.
- **Characteristics:** Includes direct quotes of statements or text. Clearly identifies specific symbols (e.g., "a swastika," "SS bolts"). Provides detailed accounts of actions or the specific content of messages/materials.
- **Example Scenario:** "The email stated 'The Holocaust is a lie'." "Graffiti included a Star of David crossed out." "They chanted 'Kill the Jews'." "The sign read 'Zionism = Nazism'."

Label Confidence Level Definitions

- **1 (Low Confidence):** Either:
  - Minimal/ambiguous objective details in `description` (correlates with Specificity 1). Hard to apply Guidelines definitively. Relies on interpretation without objective evidence. Classification is often "no" due to Lack of Information (like Example 3).
  - OR Description uses a simple 'antisemitic' label without specifics and without context suggesting conflation (Specificity 1, classified "yes" per Guideline 5 exception, like Examples 7 & 8).
- **2 (Medium Confidence):** Some objective details align with Guidelines (often Specificity 2 or 3), but minor ambiguity in applying the *rules* to those details, or the case is borderline according to the Guidelines (like Example 5). Evidence might be less explicit (like Example 6), or rule application requires some judgment.
- **3 (High Confidence):** Clear, explicit objective details in `description` (usually Specificity 3, sometimes 2) directly and unambiguously support classification according to the strict Guidelines (clear yes or clear no). Minimal ambiguity *in applying the guidelines to the described content* (Examples 1, 2, 4).

Revised Antisemitism Evaluation Guidelines (Strict Objective Interpretation)

1. Core Principles

- **Independent Evaluation**: The reported label of an incident does not determine whether it is antisemitic. Analyze only the **objective details** of the event as described in `description`. This includes the exact words used, specific symbols depicted, precise actions taken, and the immediate setting. Any commentary, interpretation, or assertion about the meaning, common usage, or effect of these elements within the `description` should be **disregarded** during evaluation, except as specifically noted in Guideline 5 regarding simple labels.

- **Intent Requires Evidence**: Antisemitism requires **evidence of hostility toward Jews as a group** *explicitly present in the objective details*, not just opposition to Israel or Zionism.
- **Context is Crucial**: Consider the immediate setting (e.g., location, accompanying objective actions or explicit language) as part of the objective details.

2. Criteria for Antisemitism

An act qualifies as antisemitic **only if the objective details in the description** meet at least one of the following:

A. Explicit Targeting of Jews (Requires explicit link in objective details)

- The act must **directly single out Jews** as a religious or ethnic group, based *only* on the explicit content in the objective details provided. This is not met solely by targeting an individual or institution known to be Jewish, unless the act itself or accompanying objective content explicitly indicates the targeting is due to Jewish identity.
    - **Qualifying Examples (Generalized)**:
        - Vandalism of a Jewish house of worship or community center *where antisemitic symbols or slurs are also present in the objective description of the vandalism*.
        - Physical or verbal assault on a person *where antisemitic language or symbols are objectively described as part of the act*.
        - Graffiti or slogans explicitly calling for harm or exclusion of Jews (e.g., "Death to Jews," "Jews not allowed") *as stated in the objective description*.
        - Targeting a Jewish individual, institution, or property with threats, harassment, or violence *where the objective description of the threat, harassment, or vandalism explicitly contains antisemitic language, symbols, or a stated antisemitic motive, or where the targeting of a Jewish individual/institution for vandalism/threats is the primary objective detail provided*.

B. Use of Classic Antisemitic Tropes or Symbols (Requires explicit presence in objective details)

- The act employs **historically rooted antisemitic stereotypes, slurs, conspiracy theories, or symbols** that are objectively present in the description.
    - **Qualifying Examples (Generalized)**:
        - Swastikas, explicitly antisemitic slurs ("kike," "shekelgrubber," etc.) *as written or stated in the description*.
        - Statements claiming Jewish control over governments, media, finance, or global events (e.g., "Jewish lobby controls X," "Jews did 9/11," "Jewish media is full of lies") *as written or stated in the description*.
        - Dehumanizing language comparing Jews to animals or demons (e.g., "Jewish roaches") *when used in an anti-Jewish context and stated in the description*.
        - Accusations of collective Jewish guilt for societal problems or historical events *explicitly stated in the description*.

C. Explicit Endorsement of Violence Against Jews (Requires explicit statement in objective details)

- The act **explicitly calls for or celebrates violence targeting Jews** (as a religious/ethnic group or Jewish civilians), not just Israeli institutions or military, based *only* on the objective details provided.
    - **Qualifying Examples (Generalized)**:
        - Slogans or statements advocating the killing of Jews worldwide ("Death to all Jews") *as stated in the description*.
        - Explicit praise for attacks specifically targeting Jewish civilians (e.g., "Happy October 7th" yelled at a Jewish person, celebrating the October 7th attack in an explicit anti-Jewish context) *as stated in the description*.
        - Distribution of material explicitly inciting violence against Jewish communities *as stated in the description*.

D. Explicit Denial or Distortion of the Holocaust (Requires explicit statement in objective details)

- The act explicitly denies the facts of the Holocaust or uses its history to promote hatred or violence against Jews, based *only* on the objective details provided.
    - **Qualifying Examples (Generalized)**:
        - Explicitly denying the systematic murder of six million Jews (e.g., "The Holocaust didn't happen," "6 Million Kikes? We Wish!").
        - Explicitly denying specific well-documented aspects of the Holocaust (e.g., "no proof of Nazi medical experiments").
        - Explicitly asserting that the Holocaust was a hoax or significantly exaggerated.
        - Explicitly using language or imagery related to the Holocaust or other genocides to promote hostility, violence, or discrimination against Jews *in a manner evident solely from the objective details* (e.g., performing a Nazi salute at a Jewish person). Equating terms like "Never again" with political actions (like "Intifada") is **not** sufficient under this criterion unless the objective details also explicitly state the connection is meant to promote violence or hatred against Jews as a group.

3. Key Distinctions (Based on Objective Details)

Not Antisemitic (Political/ Critical Speech - based on objective content)	**Antisemitic (Ethnic/Religious Hatred - based on objective content)**
Criticism of a nation-state's policies or military actions.	Attacks on Jews solely because of their identity *explicitly stated in the description.*
Advocacy for political movements opposing a government.	Conspiracy theories blaming Jews for global events *explicitly stated in the description.*
Protests featuring general slogans about liberation or resistance.	Use of slurs or symbols tied to historic persecution of Jews *explicitly stated in the description*.
Debates about nationalism, bor-	Calls for violence against Jewish people or institutions *explicitly*

Not Antisemitic (Political/ Critical Speech - based on objective content)	Antisemitic (Ethnic/Religious Hatred - based on objective content)
ders, or sovereignty.	*stated in the description*.
Strong, even inflammatory, criticism of Israel or Zionism as political entities.	Holding *all Jews* collectively responsible for actions of Israel *explicitly stated in the description*.
Calling for boycotts of Israel or Israeli goods/institutions.	Targeting Jewish individuals or businesses *solely because they are Jewish* or perceived as connected to Israel, *with antisemitic content as objective evidence or where the targeting itself is the primary objective detail*.

4. Contextual Nuances (Based on Objective Details)

- **Resistance Terminology**: Terms like "uprising" or "liberation" are not inherently antisemitic unless the objective description provides explicit evidence they are paired with explicit anti-Jewish messaging or calls for violence against Jews.
- **Historical References**: Comparisons to historical atrocities (e.g., genocide) are not automatically antisemitic unless they explicitly deny Jewish suffering or fall under Criterion 2D or are explicitly used in the objective description to promote hatred/violence against Jews.
- **Symbols**: Use of political or military symbols is not antisemitic unless the objective description provides explicit evidence they are paired with explicit anti-Jewish messaging or fall under another antisemitism criterion.

5. Handling Ambiguity (Based on Objective Details) - Revised

- **Sufficient Objective Details**: If objective details (specific words, symbols, actions) are provided, base classification solely on whether these details meet the criteria in Section 2, ignoring submitter labels or interpretations.
- **Insufficient Objective Details (Default Rule)**: If the description lacks objective details about the incident's content (e.g., uses vague terms like "hateful," "offensive," or provides no content description), default to **"not antisemitic"** due to **`Lack of Information`**. Classification requires verifiable evidence based on the guidelines.
- **Exception for Simple 'Antisemitic' Labels**: **However**, if the *only* relevant information provided about the content is a simple, direct assertion like "an antisemitic [comment/symbol/action]" **AND** there is no accompanying context in the description suggesting conflation with anti-Israel/anti-Zionist criticism (e.g., no mention of Israel, Zionism, Palestine alongside the 'antisemitic' label in a way that implies they are the *reason* for the label), then classify as **`is_antisemitic: yes`** with **`label_confidence: 1`** and **`description_specificity: 1`**. This provides a minimal "benefit of the doubt" *only* in cases of simple, unqualified, vague assertions of antisemitism where no details contradict or suggest mislabeling due to political views.

- **Conflation Suspected**: If the description pairs a vague term like "antisemitic" with anti-Israel/anti-Zionist context (e.g., "antisemitic anti-Israel statement," "antisemitic protest against Israel"), revert to the **Default Rule** (classify as "no" due to Lack of Information) because the submitter's understanding of "antisemitic" is potentially conflated and cannot be trusted without objective details.

6. Anti-Zionism vs. Antisemitism (Based on Objective Details)

- **Not Antisemitic**:
    - Criticism of Israel or Zionism as a political ideology, even if contentious or inflammatory (e.g., *"Zionism is racism"*), based *only* on the objective description.
    - Slogans linking Palestinian resistance (e.g., "Intifada") to historical Jewish suffering or concepts, *unless* the objective description explicitly includes content that advocates violence against Jews as a group, dehumanizes or scapegoats Jews in a manner consistent with classic antisemitic tropes, or explicitly falls under Criterion 2D regarding distortion of the Holocaust *through objective language*.
    - Use of Holocaust analogies to criticize Israel's policies, **unless** the objective description includes content that denies Jewish victimhood or equates Israelis/Jews to Nazis *in a manner that explicitly promotes hatred or violence against Jews* (Criterion 2D).
- **Antisemitic**:
    - Calls for violence against Jews worldwide (e.g., *"Death to all Jews"*) *explicitly stated in the description*.
    - Blaming *all Jews* for Israel's actions (e.g., *"Jewish lobby controls the U.S."*) *explicitly stated in the description*.
    - **Falling under any of the criteria in Section 2, based *only* on the objective details in the description, regardless of whether the act also involves criticism of Israel/Zionism.**

Final Instructions

Process **EACH** incident report provided in the `<data>` section below **independently**. Apply the **Antisemitism Evaluation Guidelines** strictly. Prioritize objective details. Handle vague descriptions according to the specific rules in **Guideline 5**, noting the exception for simple 'antisemitic' labels versus the default handling when details are missing or conflation is suspected. Assess **`description_specificity`**, **`is_antisemitic`**, and **`label_confidence`**. **Ignore the inflammatory nature of language unless the *described content* specifically meets criteria in Section 2.** Do **NOT** allow the content or classification of one item to influence any other. **Remember: `false_positive_reason` must be a string explanation if `is_antisemitic` is "no", and must be `null` if `is_antisemitic` is "yes".** Generate the output as a single JSON list containing

an object for each processed incident. Make sure to avoid have nested double quotes inside the json, especially in the is_antisemitic_reason

Begin processing the data below:

<data>

{data}

</data>

LLM Prompt without guidelines

Below is the prompt used when studying the biases across different LLM providers. This prompt was not used for the main classification results of our investigation.

Task: Classify Antisemitic Incidents from Descriptions

Objective: Process a list of incident reports. For each report, analyze the provided `description`- Determine the specified output fields for each incident independently.

Output Format: Provide the results as a JSON list, where each object in the list corresponds to one incident report and contains the following fields:

```
[
  {
    "description_id": "...",
    "is_antisemitic": "yes" | "no",
    "event_category": "...",
    "action_category": "...",
    "is_antisemitic_reason": "..."
  },
  {
    // ... next incident
  }
]
```

Field Definitions:

- **`description_id`**: Unique identifier (from input brackets).
- **`is_antisemitic` (yes/no)**: Based on objective `description`
- **`event_category`**: Setting/context (e.g., `Educational Institution`, `Public Area`, `Online Communication`). Choose the most appropriate.
- **`action_category`**: Type of act (e.g., `Vandalism`, `Verbal Harassment`, `Display of Symbols`, `Protest/Demonstration`). Choose the most appropriate.
- **`is_antisemitic_reason`**: Brief explanation linking description details to is_antisemitic classification.

Begin processing the data below:

<data>

{data}

</data>

Data Dictionary

Below is the data dictionary for the file adl_2016_to_2024_antisemitism_data_labelled. Columns up to description are from the ADLs dataset, and the rest are derived from this investigation.

Column Name	Data Type	Description
id	Integer	A unique identifier for each record.
year_month	Date (String)	The year and rounded down month of the incident in YYYY-MM-DD format.
date	Date (String)	The specific date of the incident in YYYY-MM-DD format.
city	String	The city where the incident occurred.
state	String	The state where the incident occurred, represented by its two-letter postal abbreviation.
ideology	String	The overarching ideology associated with the incident (if applicable).
subideology	String	A more specific subcategory of the ideology (if applicable).
type_of_attack	String	The type of attack or incident (e.g., vandalism, assault).
israel_zionism_related	String (Boolean)	Indicates whether the incident is related to Israel or Zionism (e.g., yes/no).

description	String	A detailed description of the incident.
is_antisemitic_1	String (Boolean)	Indicates if the incident is considered antisemitic (first assessment, e.g., yes/no).
is_antisemitic_2	String (Boolean)	Indicates if the incident is considered antisemitic (second assessment, e.g., yes/no).
is_antisemitic_3	String (Boolean)	Indicates if the incident is considered antisemitic (third assessment, e.g., yes/no).
false_positive_reason_1	String	Reason for classifying the incident as a false positive (first assessment, if applicable).
false_positive_reason_2	String	Reason for classifying the incident as a false positive (second assessment, if applicable).
false_positive_reason_3	String	Reason for classifying the incident as a false positive (third assessment, if applicable).
simplified_false_positive_reason_1	String	A simplified version of the first false positive reason.
simplified_false_positive_reason_2	String	A simplified version of the second false positive reason.
simplified_false_positive_reason_3	String	A simplified version of the third false positive reason.

is_antisemitic_reason_1	String	Reason for classifying the incident as antisemitic (first assessment).
is_antisemitic_reason_2	String	Reason for classifying the incident as antisemitic (second assessment).
is_antisemitic_reason_3	String	Reason for classifying the incident as antisemitic (third assessment).
label_confidence_1	Integer	Confidence level of the first antisemitic assessment (e.g., scale of 1-3).
label_confidence_2	Integer	Confidence level of the second antisemitic assessment (e.g., scale of 1-3).
label_confidence_3	Integer	Confidence level of the third antisemitic assessment (e.g., scale of 1-3).
description_specificity_1	String	Specificity of the description by the first assessment (e.g. scale of 1-3)
description_specificity_2	String	Specificity of the description by the second assessment (e.g. scale of 1-3)
description_specificity_3	String	Specificity of the description by the third assessment (e.g. scale of 1-3)
event_category_1	String	Category of the event (first assessment).
event_category_2	String	Category of the event (second assessment).

event_category_3	String	Category of the event (third assessment).
action_category_1	String	Category of action (first category).
action_category_2	String	Category of action (second category).
action_category_3	String	Category of action (third category).
contains_any_jewish_related_keywords	Integer	Indicates if the description contains any keywords related to Jewish identity. Specifically the words synagogue, kippah, rabbi.
is_antisemitic_count	Integer	A count of how many assessments classified the incident as antisemitic.
is_antisemitic_mode	String	The mode (most frequently occurring value) of the antisemitic assessments (e.g., yes/no). This corresponds to the upper bound estimate.
is_antisemitic_min	String	If at least one of the assessments says antisemitic or if contains_any_jewish_related_keywords is greater than 0. This corresponds to the lower bound estimate.
simplified_false_positive_reason_mode	String	The mode (most frequently occurring value) of the simplified false positive reasons. This corresponds to the upper bound estimate.
simplified_false_positive_reason_min	String	Contains the first available simplified false positive reason. This corresponds to the lower bound estimate.